**Homework 11: Querying data with BigQuery**

**Objective:** To query public dataset with BigQuery

**Pre-requisites:**

* Google account with access to GCP
* A project created in GCP
* BigQuery API should be enabled

**Step 1: Go to BigQuery and get the public dataset**

* Type BigQuery in the google console webpage and click on the “+” button.
* Now, enter “public dataset” in the search bar.
* Select the public dataset option.
* Now, type “NCAA” on the right pane
* Click on the “NCAA Basketball” dataset and select view dataset option.
* I have starred the datasets for my convenience, it is not mandatory.

**A screenshot of a computer

Description automatically generated**

**Step 2: Verify the “mbb\_pbp\_sr” dataset**

* Select the dataset from the left pan and click on “Preview” to see the data.
* There are other options like Schema and Details which you can check .

A screenshot of a computer

Description automatically generated

**Step 3: Query the dataset**

* Execute the below query to select the below listed fields:
  + **Query:** SELECT game\_clock, points\_scored, team\_name, event\_description FROM `bigquery-public-data.ncaa\_basketball.mbb\_pbp\_sr` WHERE season = 2016 AND home\_name = 'Panthers' AND away\_name = 'Bulls' ORDER BY timestamp DESC LIMIT 10
  + **Fields:** 
    - game\_clock
    - points\_scored
    - team\_name
    - event\_description
    - mestamp
* The below screenshot shows the query and the results.

**A screenshot of a computer

Description automatically generated**

**Description:** The above query displays the data for the season 2016 for the teams “Panthers” and “Bulls”. As the data is not wrangled, there are multiple null values for points\_scored and team\_name.

**Step 5: To update the query to include the cumulative scores for each team for a given game**

* Edit the query to give the desired output.
* **Query**:
  + SELECT game\_clock, SUM( CASE WHEN team\_name = 'Wildcats' THEN points\_scored END ) OVER(ORDER BY timestamp ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS wildcats\_score, SUM( CASE WHEN team\_name = 'Fighting Irish' THEN points\_scored END ) OVER(ORDER BY timestamp ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS fighting\_irish\_score, team\_name, event\_description FROM `bigquery-public-data.ncaa\_basketball.mbb\_pbp\_sr` WHERE season = 2014 AND home\_name = 'Wildcats' AND away\_name = 'Fighting Irish' AND points\_scored IS NOT NULL ORDER BY timestamp DESC LIMIT 10;
* The below screenshot shows the query with the output.

**A screenshot of a computer

Description automatically generated**

**Description:** With this query we can get the scores for the entire game for each event of the game.

**Clauses Used:**

* **SUM:** gets the sum of the fields specified
* **CASE:**  is a conditional operator, in which we can define the conditions and the required operation when the given condition is satisfied.
* **Over:** references a group of rows
* **Order by:** used to sort the data in ascending or descending order by a given field

**Conclusion:** We used a public dataset to query the data and found out that the dataset is not cleaned hence, there were multiple null values. Then we modified our query to get data for cumulative sum of scores for each team for the game.